

ICE: Information Credibility Evaluation on Social Media via Representation Learning

Qiang Liu, Shu Wu, *Member, IEEE*, Feng Yu, Liang Wang, *Senior Member, IEEE*, and Tieniu Tan, *Fellow, IEEE*

Abstract—With the rapid growth of social media, rumors are also spreading widely on social media and bring harm to people's daily life. Nowadays, information credibility evaluation has drawn attention from academic and industrial communities. Current methods mainly focus on feature engineering and achieve some success. However, feature engineering based methods require a lot of labor and cannot fully reveal the underlying relations among data. In our viewpoint, the key elements of user behaviors for evaluating credibility are concluded as who, what, when, and how. These existing methods cannot model the correlation among different key elements during the spreading of microblogs. In this paper, we propose a novel representation learning method, Information Credibility Evaluation (ICE), to learn representations of information credibility on social media. In ICE, latent representations are learnt for modeling user credibility, behavior types, temporal properties, and comment attitudes. The aggregation of these factors in the microblog spreading process yields the representation of a users behavior, and the aggregation of these dynamic representations generates the credibility representation of an event spreading on social media. Moreover, a pairwise learning method is applied to maximize the credibility difference between rumors and non-rumors. To evaluate the performance of ICE, we conduct experiments on a Sina Weibo data set, and the experimental results show that our ICE model outperforms the state-of-the-art methods.

Index Terms—information credibility evaluation, rumor detection, social media, representation learning.

I. INTRODUCTION

WITH the rapid growth of social media, such as Facebook, Twitter, and Sina Weibo, people are able to share information and express their attitudes publicly. Social media brings great convenience to users, and information can be spread more rapidly and widely nowadays. At the same time, rumors can also be spread on the Internet more easily and viewed by more people. A rumor is an unverified and instrumentally relevant statement of information spreading among people [5]. Rumors bring significant harm to daily life, social harmony, or even public security. With the growth of the Internet and social media, such harm will also grow greater. For instance, as the loss of MH370 has drawn worldwide attention, a great amount of rumors has spread on social media, e.g., MH370 has landed in China,¹ the loss of MH370 is

caused by terrorists,² and Russian jets are related to the loss of MH370.³ These rumors about MH370 mislead public attitudes to a wrong direction and delay the search of MH370. Up to October 10, 2015, on the biggest Chinese microblog website Sina Weibo,⁴ 28,454 rumors have been reported and collected in its misinformation management center.⁵ Accordingly, it is crucial to evaluate information credibility and to detect rumors on social media.

Nowadays, to automatically evaluate information credibility on social media, some methods have been proposed. Existing methods are mainly based on feature engineering, i.e., methods with handcrafted features. Most of the methods are based on content information and source credibility at the microblog level [3][27][10] or event (containing a group of microblogs) level [16][42][22]. Some research also studies the aggregation of credibility from the microblog level to the event level [14]. On the contrary, considering dynamic information, some work designs temporal features based on prorogation properties over time [16] or trains a model with features generated from different time periods [22]. Some works also take usage of users' feedbacks (comments and attitudes) to evaluate credibility [8][29]. The Enquiry Post (EP) model [42] takes out signal tweets, which indicates users' suspicious attitudes for detecting rumors and achieves satisfactory performance.

Although these methods have been widely used, they have several drawbacks. First, these methods based on feature engineering require great labor for designing features [3]. Moreover, a rough fusion resting on the statistical summation of these feature values is not competent to model elaborate interactions among different features on social media for information credibility evaluation. For instance, there are two combinations: (1) "a user with high credibility posted a microblog" and "a user with low credibility reposted a microblog" and (2) "a user with low credibility posted a microblog" and "a user with high credibility reposted a microblog". The values of the corresponding features, such as user credibility (high or low) and behavior type (post or repost) from the above two combinations, are equal by statistical summation. Therefore, a rough fusion rested on statistical summation cannot tell these two combinations apart. Intuitively, the former combination is more like the style of non-rumor and the later combination is more like the

The authors are with the University of Chinese Academy of Sciences (UCAS) and the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, 100190, China. E-mail: {qiang.liu, shu.wu, feng.yu, wangliang, tnt}@nlpr.ia.ac.cn.

¹<http://www.fireandreamitchell.com/2014/03/07/rumor-malaysia-airlines-mh370-landed-china/>

²<http://www.csmonitor.com/World/Asia-Pacific/2014/0310/Malaysia-Airlines-flight-MH370-China-plays-down-terrorism-theories-video>

³<http://www.inquisitr.com/1689765/malaysia-airlines-flight-mh370-russian-jets-in-baltic-may-hold-clue-to-how-flight-370-vanished/>

⁴<http://weibo.com>

⁵<http://service.account.weibo.com/?type=5&status=4>

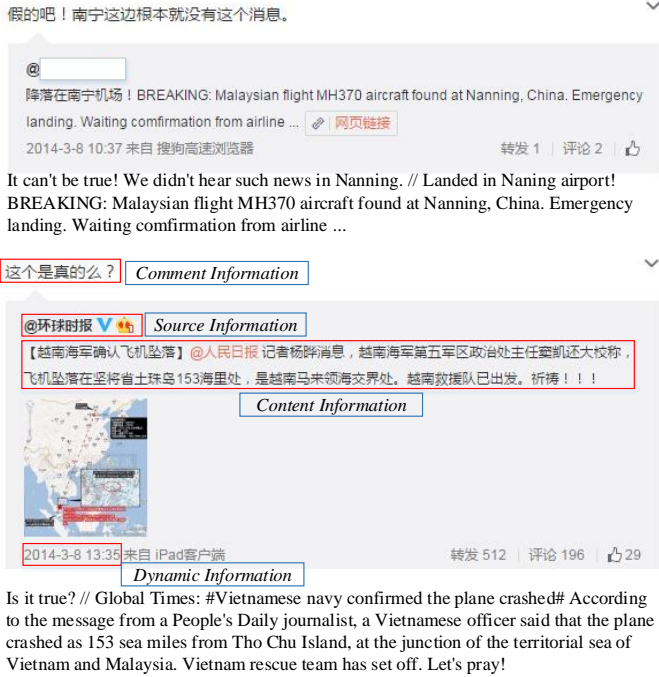


Fig. 1. Two rumor examples about MH370 and their repostings on Sina Weibo. Their corresponding English translations are listed below them.

style of rumor. On the contrary, methods based on feature engineering cannot model some real-world scenarios from a joint perspective, i.e., who did what at when and how others reacted. Those methods treat factors (who, what, when, how) as separate features and can extract simple compound features, such as a user with low credibility tended to post a rumor. There are complicated compound features, such as a user with low credibility posted a rumor at early stage of spreading receiving suspicion comments and a user with high credibility reposted a rumor at medium term of spreading receiving identification comments. The analysis of those complicated compound features from statistical summation requires enumerating all possible complicated compound features, which results in the explosion of time complexity and problem of data sparsity. What's more, a complicated compound feature may include some simple compound features, so the statistical summation of those enumerated compound features cannot truly reflect the distribution of semantic information, because a simple compound feature combined with another simple compound may generate a reverse meaning, such as "a user with high credibility" may be a non-rumor style and "a user with high credibility reposted a rumor at medium term of spreading" may be a rumor style. As a consequence, it is better to model elaborate interactions among different features to obtain an overall and joint understanding of complicated behaviors on social media.

In this paper, we evaluate the credibility of information about events on social media. Usually, each event contains several microblogs posted and reposted by users. To identify whether an event is a rumor or not, we first investigate microblogs of events on social media. Figure 1 shows some examples of rumors on Sina Weibo with extracted source

information, content information, temporal information, and comment information. According to this information on social media, we conclude four key factors for evaluating information credibility on social media:

(1) "Who" means source credibility or user credibility. Generally speaking, a source usually means a user. Normally, the higher the credibility of a user, the higher the credibility of information it creates [3]. However, some studies [13] point out that a great amount of users with high credibility on social media would repost and share misinformation unintentionally. As shown in the example in Figure 1, even a regular media (usually with high credibility), Global Times, would post unverified news about MH370. Therefore, it is not reliable to model user credibility information alone for information credibility evaluation.

(2) "What" denotes behavior types. Usually, there are two types of behaviors for users, i.e., posting and reposting. Compared to reposts, an original post indicates that the microblog is more original and relatively more important for evaluating credibility. For non-rumors, original microblogs are usually posted by users with high credibility. For rumors, original microblogs are posted by users with low credibility, whereas users with high credibility may repost the microblogs.

(3) "When" refers to temporal properties that describe the spread process of a microblog post. As shown in Figure 2(a), temporal properties are usually different between rumors and non-rumors. Compared to rumors, most non-rumor microblogs tended to be posted or reposted at the beginning and vanish very fast. Maybe those plain truths will become less and less attractive as time goes on. However, rumors usually draw comparatively sustained attention. Moreover, the spreading curve of rumors may have multiple peaks. There may be some rumormongers promoting the spreading of rumors. In addition, for non-rumors, original microblogs at the beginning event are usually posted by users with high credibility, whereas, for rumors, original microblogs are usually posted by users with low credibility and then reposted by other users including those with high credibility.

(4) "How" denotes comments and attitudes towards corresponding microblogs. Users on social media can express their attitudes and collective intelligence can be gathered to help us evaluate the credibility of information [8]. Comments reveal the users suspicion or identification attitudes towards microblogs. As shown in Figure 2(b), rumors usually receive more suspicion comments, which is extremely helpful for detecting rumors.

The aggregation of these key factors makes the joint perspective of a microblog and helps evaluate credibility. However, conventional feature engineering-based methods consider these factors as separate features and roughly summarize them, which cannot model the interactions among the key factors. Accordingly, we plan to learn representations to obtain an overall and joint understanding of complicated and dynamic behaviors in information spreading and evaluate the credibility, which are shaped by modeling elaborate interactions among different features. To be specific, we model semantic operations among different features and form an overall representation of each microblog post. Recently, representa-

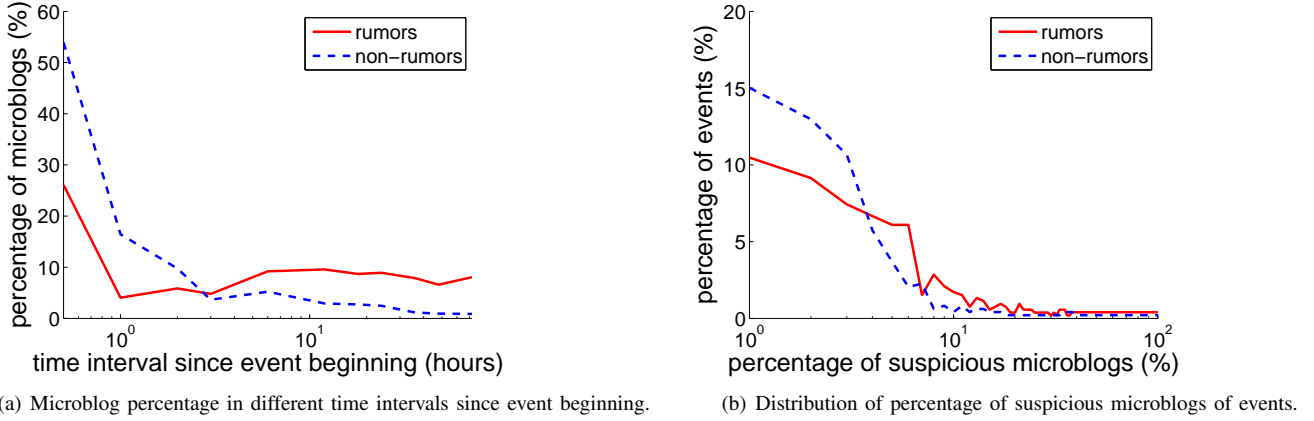


Fig. 2. Analysis of data distribution difference between rumors and non-rumors on the Weibo data set.

tion learning [1] is showing a promising performance in a variety of applications, such as word embedding [23][24][25], network embedding [2][9][26][32], and user representations [6][7][18][19].

To the end, we propose a novel ICE model that learns the joint representations of key factors of microblogs and further evaluates the credibility of events. In ICE, each user is represented as a vector according to his or her personal features, indicating the credibility information of the user (“who”). For the sake of modeling elaborate interactions among different features, other features such as “what”, “how”, and “when” are represented as operating matrices [19]. For instance, behavior types (post or repost) are modeled as latent operating matrices indicating the properties of different behaviors (“what”) and we incorporate matrix representations for time intervals since the beginning of spreading of an event’s information to capture the temporal properties of behaviors (“when”). Moreover, the attitudes of comments (suspicious or not) are also modeled as latent operating matrices indicating collective attitudes (“how”). These operating matrices model semantic operations of one feature on another. Consequently, based on the representations of users, the representations of dynamic and complicated behaviors can be obtained through multiplication with all operating matrices of varied features. Each representation of a dynamic behavior can also be viewed as a representation of a corresponding microblog about a specific event. After aggregating all the microblog representations during information spreading, we can generate the credibility representation of the event. Then, we apply a pairwise learning method to enlarge the credibility difference between rumors and non-rumors for a better and fast learning of parameters. We crawl a data set from Sina Weibo, and experiments show that our model achieves better performance compared to state-of-the-art methods.

The main contributions of this work are listed as follows:

- We introduce a representation learning method for information credibility evaluation. The proposed method captures elaborate interactions among the key factors of microblogs during information spreading through learning operating matrices, which model abundant semantic

operations among varied features.

- ICE learns latent representations for user credibility, behavior types, temporal properties and attitudes of comments. Based on these representations, ICE generates overall credibility representations of information and presents a novel perspective on information credibility evaluation.
- Experiments conducted on a real-world data set show that ICE is effective and clearly outperforms the state-of-the-art methods.

The rest of the paper is organized as follows. In Section 2, we review some related work on truth discovery, credibility evaluation, and representation learning. In Section 3, we introduce our data set and give some analysis. Section 4 details our ICE model. In Section 5, we report experimental results on the Weibo data set and compare them to several state-of-the-art methods. In Section 6, we present a real-time information credibility evaluation system that we constructed based on our proposed model. Section 7 concludes our work and discusses future research.

II. RELATED WORK

In this section, we review some related works, including credibility evaluation on social media, representation learning and truth discovery.

A. Credibility Evaluation on Social Media

Recently, some works have been proposed to automatically evaluate the information credibility and detect rumors on social media. Most of the methods are based on artificial features. Some of them evaluate the credibility of a single microblog [3][27] or a single image [10]. Some of them evaluate information credibility at the event level to distinguish whether an event is a rumor or a non-rumor [11][31][16][42][22], where each event consists of several microblogs. News Credibility Propagation (NewsCP) [14] studies how to aggregate credibility from the microblog level to the event level and presents a graph optimization method, which has further incorporated conflict viewpoints in the model [15]. Some works detect

TABLE I
DETAILS OF THE WEIBO DATA SET.

#events	#rumors	#non-rumors	#microblogs	#postings	#repostings	#users
936	500	436	630,665	98,429	532,236	321,246

rumors based on the dynamic properties. For instance, the Periodic External Shocks (PES) model [16] uses ordinary structural features and user features and designs temporal features according to the properties of information spreading over time. The Dynamic Series-Time Structure (DSTS) [22] generates content-, user-, and diffusion-based features in different time periods during information spreading and uses all these features to train a model. Some works also take usage of users feedbacks to evaluate credibility [8][29]. The EP model [42] extracts signal tweets that indicate users suspicious attitudes for detecting rumors and achieves satisfactory performance. The main drawback of these feature engineering-based models lies in that they require great labor for designing a great many features and cannot reveal underlying relations among these features. Moreover, these methods have difficulty in modeling elaborate interactions among different factors during information spreading.

B. Representation Learning

Nowadays, representation learning [1] has been extensively studied in different areas. In natural language processing, learning embeddings [25] is a hot topic, where recurrent neural networks [23][24] are widely applied. In web mining, learning network embedding has drawn great attention for studying node classification [12] or information diffusion [2]. Recently, network embedding models have incorporated random walk [26][9] and second-order connection in the representation learning methods [32]. Meanwhile, representation models are also playing a role for modeling user behaviors. Contextual Operating Tensor (COT) [19][36] and CARS2 [30] study context-aware user representations for recommendation. Hierarchical Interaction Representation (HIR) [18] studies joint representations of entities, e.g., users, items and contexts, to model their interaction. Some works [6][37][41] utilize deep neural networks for better user modeling. Convolutional Click Prediction Model (CCPM) [21] applies convolutional neural networks in predicting clicking behaviors of users. Hierarchical Representation Model (HRM) [34] and Dynamic Recurrent Basket Model (DREAM)[40] learn the representation of behaviors of a user in a short period for better recommendation. These methods achieve the state-of-the-art performance in different areas, and give us inspiration for learning representations of dynamic behaviors to evaluate information credibility.

Nowadays, representation learning [1] has been extensively studied in different areas. In natural language processing, learning embedding [25] is a hot topic, where recurrent neural networks [23][24] are widely applied. In web mining, learning network embedding has drawn great attention for studying node classification [12] or information diffusion [2]. Recently, network embedding models have incorporated random walk

[26][9] and second-order connection in the representation learning methods [32]. Meanwhile, representation models are also playing a role for modeling user behaviors. The contextual operating tensor (COT) [19][36] and CARS2 [30] study context-aware user representations for recommendation. Hierarchical Interaction Representation (HIR) [18] studies joint representations of entities, e.g., users, items, and contexts, to model their interaction. Some works [6][37][41] use deep neural networks for better user modeling. The convolutional click prediction model (CCPM) [21] applies convolutional neural networks in predicting clicking behaviors of users. The hierarchical representation model (HRM) [34] and the dynamic recurrent basket model (DREAM) [40] learn the representation of behaviors of a user in a short period for better recommendation. These methods achieve state-of-the-art performance in different areas and give inspiration for learning representations of dynamic behaviors to evaluate information credibility.

C. Truth Discovery

Truth discovery refers to the problem of finding the truth with conflicting information, which has been first addressed in [38]. It can be viewed as some kind of information credibility evaluation. Mainly based on the source credibility information, truth discovery evaluates the credibility via aggregating from different sources. Truth discovery methods are usually based on Bayesian algorithms or graph learning algorithms on stock data or flight data [17][33]. And Semi-Supervised Truth Discovery (SSTF) [39] studies the problem with semi-supervised graph learning with a small set of ground truth data to help evaluating credibility. Truth discovery is an unsupervised or semi-supervised method to find the truth with conflicting information and make an evaluation of information credibility [17]. Truth discovery is mainly based on the evaluated credibility aggregated from different information sources, usually referring to users who release information. And it is not capable to make the most of various kinds of information, such as time properties and comment attitudes, which are abundant in complex online social media scenario. Therefore, truth discovery is suitable for ideal situations with constrained topics, such as price prediction and flight arrival prediction. They are hard to be applied in complex online social media.

III. DATA

In this section, we introduce our data set. Considering that there is a lack of public rumor data sets, we collected a microblog data set containing rumors and non-rumors from Sina Weibo, which is the biggest social media in China.

To crawl rumors, we collected some rumor seeds, i.e., some microblogs containing rumors that have been reported,

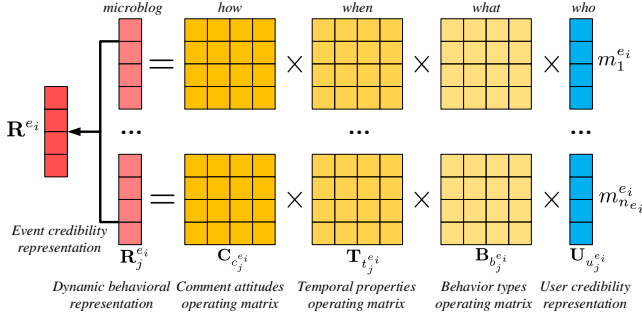


Fig. 3. Overview of the representation learning procedure in the proposed ICE model.

from misinformation management center of Sina Weibo. We extracted keywords from these rumor seeds and retrieved rumor microblogs with these keywords. Then, we identified the starting point of a rumor, i.e., the first microblog about the rumor, and collected all the following microblogs. For each microblog, we collected its reposting information, commenting information, and the corresponding users profile. To crawl non-rumors, we collected some hot topics on Sina Weibo and used the same strategy as for rumors to crawl corresponding information about the non-rumors.

As shown in Table I, we collected 936 events containing 500 rumors and 436 non-rumors. Each event consists of several microblogs (postings or repostings), and the average number is about 673. The total number of microblogs is 630,363, including 98,429 postings and 532,236 repostings. Each microblog has its posting time. The posting time of the first microblog about an event is set as the beginning of the event. The data set contains 321,246 users. The personal profile of a user includes gender, verified or not, number of followers, number of followees, and number of microblogs.

Moreover, considering that it is necessary to mine suspicion and identification attitudes towards microblogs from comments, we need to annotate each microblog suspicious or not. For there is no proper corpus for training a classifier about suspicion, we used an unsupervised method to identify suspicious attitudes towards microblogs. We built up a list of suspicion words and distinguished a microblog according to whether those suspicion words appear in the microblog. We first found several typical suspicion words then train word2vec⁶ [25] on our data set and found dozens of words similar with the typical suspicion words according to their embedding distance. Finally, we built up a word list with about 100 suspicion words and annotated all the microblogs suspicious or not in our Weibo data set.

Based on the data set, we also investigated the data distribution difference between rumors and non-rumors, which is shown in Figure 2, i.e., distribution of percentage of microblogs with time illustrated in Figure 2(a) and distribution of percentage of events with the percentage of suspicious microblogs in one event shown in Figure 2(b).

IV. THE ICE MODEL

In this section, we first formulate the problem. Then, we detail the proposed ICE model. Finally, we present the pairwise learning procedure for the ICE model.

A. Problem Formulation

The problem we studied in this paper can be formulated in math as follows. Suppose a set of events are denoted as $E = \{e_1, e_2, \dots, e_n\}$ and s_{e_i} is the credibility score of the corresponding event e_i . $s_{e_i} = 0$ means event e_i is a rumor and $s_{e_i} = 1$ means event e_i is a non-rumor. The microblogs of the event e_i can be denoted as $M^{e_i} = \{m_1^{e_i}, m_2^{e_i}, \dots, m_{n_{e_i}}^{e_i}\}$, where n_{e_i} is the number of microblogs of this event. All microblog sets can be written as $M = \{M^{e_1}, M^{e_2}, \dots, M^{e_n}\}$. Each microblog $m_j^{e_i}$ consists of four elements “who”, “what”, “how”, and “when”, which are denoted as $u_j^{e_i}$, $b_j^{e_i}$, $c_j^{e_i}$ and $t_j^{e_i}$. $u_j^{e_i}$ is the corresponding user of the microblog, $b_j^{e_i}$ denotes the behavior type (posting or reposting), $c_j^{e_i}$ describes the user’s comments and attitudes (suspicious or not), and $t_j^{e_i}$ denotes the time interval since the beginning of an event. In this work, our task is to evaluate the credibility of an event on social media.

B. Proposed Model

Here, we detail the representation learning procedure of the ICE model. Based on handcrafted features that indicate global-wise statistics, conventional methods have difficulty in modeling the correlation among different key elements in information spreading. Thus, we need to model their joint representations and yield their joint characteristics. It is necessary for a model based on user credibility (who), behavior types (what), comment attitudes (how) and dynamic properties (when).

We first start with user information and behavior information. User information tells us the properties and credibility of a user. Behavior information tells us the behavior type, i.e., posting or reposting. Moreover, the combination of these two information shows “who did what”. Mathematically, for the j -th microblog $m_j^{e_i}$ of the event e_i , the representation of this microblog with the user $u_j^{e_i}$ and the behavior $b_j^{e_i}$ can be written as

$$R_j^{e_i} = B_j^{e_i} U_j^{e_i}, \quad (1)$$

where $U_j^{e_i} \in \mathbb{R}^d$ is the vector representation of user $u_j^{e_i}$, $B_j^{e_i} \in \mathbb{R}^{d \times d}$ is the matrix representation of behavior $b_j^{e_i}$, and d denotes the dimensionality of representations.

Additionally, users may express their attitudes in the comments. These attitudes contain the knowledge and life experience of users and can be used to distinguish rumors from non-rumors. As shown in Figure 2(b), rumors often receive more suspicious comments than non-rumors. Incorporating the representation of comment attitude $c_j^{e_i}$ of microblog $m_j^{e_i}$, Equation 1 can be written as

$$R_j^{e_i} = C_j^{e_i} B_j^{e_i} U_j^{e_i}, \quad (2)$$

⁶<https://code.google.com/p/word2vec/>

where $\mathbf{C}_j^{e_i} \in \mathbb{R}^{d \times d}$ is the matrix representation of the comment $c_j^{e_i}$. Now, this equation can reveal the joint representation of “who did what under how”.

Moreover, Figure 2(a) illustrates the difference between dynamic properties of rumors and non-rumors. It shows that time interval information is a significant factor for evaluating information credibility and should be modeled jointly with user behaviors. For instance, time interval $t_j^{e_i}$ of microblog $m_j^{e_i}$ means that the microblog appears from the beginning of e_i . Incorporating time interval $t_j^{e_i}$ in Equation 2, the representation of microblog $m_j^{e_i}$ can be written as

$$\mathbf{R}_j^{e_i} = \mathbf{T}_j^{e_i} \mathbf{C}_j^{e_i} \mathbf{B}_j^{e_i} \mathbf{U}_j^{e_i}, \quad (3)$$

where $\mathbf{T}_j^{e_i} \in \mathbb{R}^{d \times d}$ is the matrix representation of time interval $t_j^{e_i}$. Now, this equation can reveal the joint representation of “who did what under how at when”.

Right now, we generate the representation $\mathbf{R}_j^{e_i}$ of the microblog $m_j^{e_i}$, which can capture the joint properties of four key elements. Because each event consists of several microblogs, we need to aggregate representations of microblogs to generate the final credibility representation of the event. Using the average calculation, the representation of the event e_i can be generated as

$$\begin{aligned} \mathbf{R}^{e_i} &= \frac{1}{n_{e_i}} \sum_{m_j^{e_i} \in M^{e_i}} \mathbf{R}_j^{e_i} \\ &= \frac{1}{n_{e_i}} \sum_{m_j^{e_i} \in M^{e_i}} \mathbf{T}_j^{e_i} \mathbf{C}_j^{e_i} \mathbf{B}_j^{e_i} \mathbf{U}_j^{e_i}. \end{aligned} \quad (4)$$

Then, we can predict whether an event e_i is a rumor or not using

$$y^{e_i} = \mathbf{W}^T \mathbf{R}^{e_i}, \quad (5)$$

where $\mathbf{W} \in \mathbb{R}^d$ is linear weights of the prediction function. A larger value of y^{e_i} indicates higher credibility of e_i .

C. User Representation Generation

For learning user representations, i.e., “who”, in the ICE model, it would be desirable if we can learn a distinct latent vector for each user to capture his or her properties and credibility. However, according to Table I, each user (re)posts only two microblogs in average, which can not bring enough information to directly learn a latent representation for each user.

Instead, we can learn embeddings of rich features for users. These features contained in the Weibo data set are gender, number of followers, number of followees, numbers of microblogs, and verified or not. Then, users can be shaped based on the above features. For user u , we have a feature vector $\mathbf{F}_u \in \mathbb{R}^f$ which is constructed as

$$\mathbf{F}_u = [\mathbf{F}_u^{\text{gender}}, \mathbf{F}_u^{\text{followers}}, \mathbf{F}_u^{\text{followees}}, \mathbf{F}_u^{\text{microblogs}}, \mathbf{F}_u^{\text{verified}}]^T$$

Both $\mathbf{F}_u^{\text{gender}}$ and $\mathbf{F}_u^{\text{verified}}$ have two bits. $\mathbf{F}_u^{\text{gender}}(1) = 1$ denotes that the gender is male, and $\mathbf{F}_u^{\text{gender}}(2) = 1$ denotes that the gender is female. $\mathbf{F}_u^{\text{verified}}(1) = 1$ means that the user is verified, and $\mathbf{F}_u^{\text{verified}}(2) = 1$ otherwise. For the numbers of followers, followees, and microblogs, it is hard to learn an

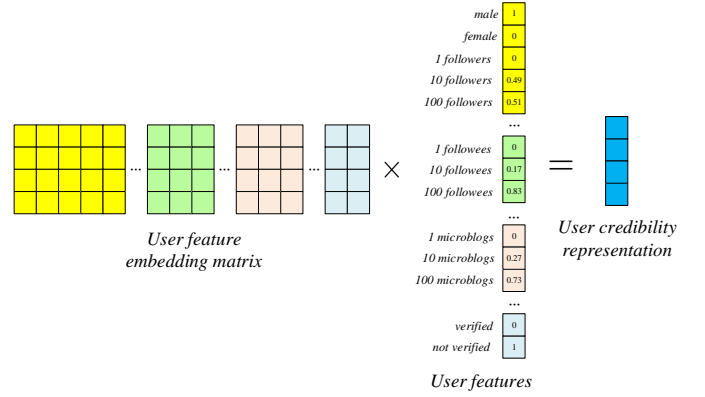


Fig. 4. An example of generating user representation. The user’s features are {male, 32 followers, 68 followees, 54 microblogs, not verified}.

embedding for each distinct value. Therefore, we partition the values into discrete bins according to a \log_{10} distribution. If a user u has v_u followers, the corresponding features can be constructed as

$$\mathbf{F}_u^{\text{followers}}(i) = \begin{cases} U(\log_{10}^{v_u}) - \log_{10}^{v_u}, & i = L(\log_{10}^{v_u}) + 1 \\ \log_{10}^{v_u} - L(\log_{10}^{v_u}), & i = U(\log_{10}^{v_u}) + 1 \\ 0, & i = \text{others} \end{cases}$$

where $U(\log_{10}^{v_u})$ and $L(\log_{10}^{v_u})$ denote the upper and lower bounds of $\log_{10}^{v_u}$ respectively. Meanwhile, $\mathbf{F}_u^{\text{followees}}$ and $\mathbf{F}_u^{\text{microblogs}}$ can be constructed in the same way. Figure 4 illustrates an example of generating user representation. In the example, suppose $v_u = 32$, then $\log_{10}^{32} = 1.51$, the corresponding upper and lower bounds are 2 and 1. $\mathbf{F}_u^{\text{followers}}$ can be computed as

$$\mathbf{F}_u^{\text{followers}}(2) = 2 - 1.51 = 0.49,$$

$$\mathbf{F}_u^{\text{followers}}(3) = 1.51 - 1 = 0.51,$$

and other bits will be set to be 0.

Then, based on feature vector \mathbf{F}_u , we can generate the user representation as

$$\mathbf{U}_u = \mathbf{S} \mathbf{F}_u, \quad (6)$$

where $\mathbf{S} \in \mathbb{R}^{d \times f}$ is the feature embedding matrix.

D. Nonlinear Interpolation for Generating Time-Specific Matrices

In ICE, we use time-specific matrices to capture the properties of users’ dynamic behaviors, i.e., “when”, in the information spreading. However, if we learn a distinct matrix for each possible continuous time interval, the ICE model will face the data sparsity problem. Therefore, as in [20], we use a similar strategy for generating time-specific matrices. We partition the time interval into discrete time bins. Considering the power law distribution of dynamic behaviors shown in Figure 2(a), it is not plausible if we partition the time interval equally. Instead, our partition conforms to a \log_2 distribution. Only the matrices of the upper and lower bounds of the corresponding bins are learned in our model. For time intervals in a time

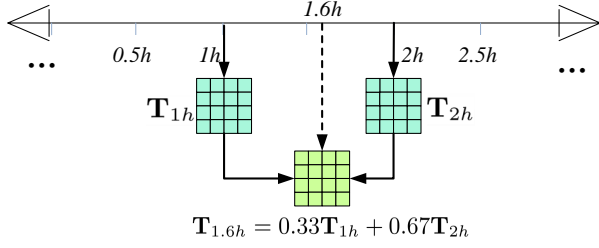


Fig. 5. An example of generating time-specific matrix. The time interval in this example is $1.6h$. Via a nonlinear interpolation of $1h$ and $2h$, the corresponding time-specific matrix can be generated as $\mathbf{T}_{1.6h} = 0.33\mathbf{T}_{1h} + 0.67\mathbf{T}_{2h}$.

bin, their transition matrices can be calculated via a nonlinear interpolation.

Mathematically, the time-specific matrix \mathbf{T}_t for time interval t can be calculated as

$$\mathbf{T}_t = \frac{(U(\log_2^t) - \log_2^t)\mathbf{T}_{2^{L(\log_2^t)}} + (\log_2^t - L(\log_2^t))\mathbf{T}_{2^{U(\log_2^t)}}}{U(\log_2^t) - L(\log_2^t)},$$

where $U(\log_2^t)$ and $L(\log_2^t)$ denote the upper bound and lower bounds of \log_2^t respectively. An example is shown in Figure 5, where $t = 1.6h$, $\log_2^{1.6h} = 0.67$, the corresponding upper and lower bounds will be 1 and 0, respectively; then, $\mathbf{T}_{1.6h}$ can be computed as

$$\begin{aligned} \mathbf{T}_{1.6h} &= \frac{(1 - 0.67)\mathbf{T}_{1h} + (0.67 - 0)\mathbf{T}_{2h}}{1 - 0} \\ &= 0.33\mathbf{T}_{1h} + 0.67\mathbf{T}_{2h} \end{aligned}$$

Such an interpolation method can solve the problem of learning matrices for continuous values in the ICE model and provide a solution for modeling the dynamic behaviors of users.

E. Pair-wise Learning

Here, we introduce the parameter estimation process of ICE with a pair-wise learning method and calculate the complexity of the algorithm.

Because rumors are often hard to collect for training credibility evaluation models, we apply a pair-wise learning method to enlarge the number of training instances. Similar to [28], our basic assumption is that the credibility of a non-rumor is larger than that of a rumor. In ICE, we can maximize the credibility difference between rumors and non-rumors. Accordingly, we should maximize the following probability:

$$p(e_n \succ e_r) = g(y^{e_n} - y^{e_r}),$$

where e_n denotes a non-rumor, e_r denotes a rumor, and $g(x)$ is a nonlinear function that is selected as:

$$g(x) = \frac{1}{1 + e^{-x}}.$$

Incorporating the negative log likelihood, for the whole data set, we can minimize the following objective function equivalently:

$$J = \sum_{\{e_n, e_r\} \in E, l_{e_n}=1, l_{e_r}=0} \ln(1 + e^{-\mathbf{W}^T(\mathbf{R}^{e_n} - \mathbf{R}^{e_r})}) + \frac{\lambda}{2} \|\Theta\|^2,$$

where $\Theta = \{\mathbf{U}, \mathbf{B}, \mathbf{C}, \mathbf{T}, \mathbf{W}\}$ denotes all the parameters to be estimated and λ is a parameter to control the power of regularization. The derivations of J with respect to \mathbf{W} , \mathbf{R}^{e_n} and \mathbf{R}^{e_r} can be calculated as

$$\begin{aligned} \frac{\partial J}{\partial \mathbf{W}} &= \sum_{e_n, e_r \in E, l_{e_n}=1, l_{e_r}=0} \frac{(\mathbf{R}^{e_r} - \mathbf{R}^{e_n})l(e_n, e_r)}{1 + l(e_n, e_r)} + \lambda \mathbf{W}, \\ \frac{\partial J}{\partial \mathbf{R}^{e_n}} &= - \sum_{e_r \in E, l_{e_r}=0} \frac{\mathbf{W}l(e_n, e_r)}{1 + l(e_n, e_r)}, \\ \frac{\partial J}{\partial \mathbf{R}^{e_r}} &= \sum_{e_n \in E, l_{e_n}=1} \frac{\mathbf{W}l(e_n, e_r)}{1 + l(e_n, e_r)}. \end{aligned}$$

where

$$l(e_n, e_r) = e^{-\mathbf{W}^T(\mathbf{R}^{e_n} - \mathbf{R}^{e_r})}.$$

After calculating the derivation $\partial J / \partial \mathbf{R}^{e_i}$ of event representation \mathbf{R}^{e_i} , the corresponding gradients all the parameters can be calculated as

$$\begin{aligned} \frac{\partial J}{\partial \mathbf{T}_j^{e_i}} &= \frac{1}{n_{e_i}} \frac{\partial J}{\partial \mathbf{R}^{e_i}} (\mathbf{C}_j^{e_i} \mathbf{B}_j^{e_i} \mathbf{U}_j^{e_i})^T, \\ \frac{\partial J}{\partial \mathbf{C}_j^{e_i}} &= \frac{1}{n_{e_i}} (\mathbf{T}_j^{e_i})^T \frac{\partial J}{\partial \mathbf{R}^{e_i}} (\mathbf{B}_j^{e_i} \mathbf{U}_j^{e_i})^T, \\ \frac{\partial J}{\partial \mathbf{B}_j^{e_i}} &= \frac{1}{n_{e_i}} (\mathbf{T}_j^{e_i} \mathbf{C}_j^{e_i})^T \frac{\partial J}{\partial \mathbf{R}^{e_i}} (\mathbf{U}_j^{e_i})^T, \\ \frac{\partial J}{\partial \mathbf{U}_j^{e_i}} &= \frac{1}{n_{e_i}} (\mathbf{T}_j^{e_i} \mathbf{C}_j^{e_i} \mathbf{B}_j^{e_i})^T \frac{\partial J}{\partial \mathbf{R}^{e_i}}. \end{aligned}$$

Then,

$$\begin{aligned} \frac{\partial J}{\partial \mathbf{T}_t} &= \sum_{e_i \in E} \sum_{m_j^{e_i} \in M^{e_i}, t_j^{e_i}=t} \frac{\partial J}{\partial \mathbf{T}_j^{e_i}} + \lambda \mathbf{T}_t, \\ \frac{\partial J}{\partial \mathbf{C}_c} &= \sum_{e_i \in E} \sum_{m_j^{e_i} \in M^{e_i}, c_j^{e_i}=c} \frac{\partial J}{\partial \mathbf{C}_j^{e_i}} + \lambda \mathbf{C}_c, \\ \frac{\partial J}{\partial \mathbf{B}_b} &= \sum_{e_i \in E} \sum_{m_j^{e_i} \in M^{e_i}, b_j^{e_i}=b} \frac{\partial J}{\partial \mathbf{B}_j^{e_i}} + \lambda \mathbf{B}_b, \\ \frac{\partial J}{\partial \mathbf{U}_u} &= \sum_{e_i \in E} \sum_{m_j^{e_i} \in M^{e_i}, u_j^{e_i}=u} \frac{\partial J}{\partial \mathbf{U}_j^{e_i}} + \lambda \mathbf{U}_u. \end{aligned}$$

After all the gradients are calculated, we can employ gradient descent to estimate the model parameters. This process can be repeated iteratively until convergence.

Based on the above calculation, now we analyze the corresponding time complexity and suppose we totally have n event with m microblogs. During the training procedure, in each iteration, the time complexities of updating T , C , B and U are $O(d^2 \times m)$ respectively. And the time complexity of updating W is $O(d \times n)$. So, the total time complexity is $O[4d^2 \times m + d \times n]$. Since m is usually much larger than n and d is a constant, the time complexity is approximately equal to $O(m)$. During the testing procedure, the time complexity is $O(d^2 \times m + d \times n)$. It is also approximately equal to $O(m)$. This indicates that both training and testing time complexities grow linearly with size of the dataset, and ICE has potential to scale up to large-scale data.

TABLE II
PERFORMANCE COMPARISON EVALUATED WITH DIMENSIONALITY $d = 8$.

Methods	Accuracy	Rumors			Non-rumors		
		Precision	Recall	F1-score	Precision	Recall	F1-score
NewsCP-Content	0.608	0.524	0.899	0.662	0.531	0.782	0.633
NewsCP-Social	0.618	0.617	0.929	0.742	0.556	0.793	0.654
NewsCP	0.758	0.741	0.808	0.773	0.728	0.770	0.749
EP	0.812	0.795	0.899	0.844	0.802	0.793	0.798
EP+Content	0.823	0.809	0.899	0.852	0.868	0.759	0.810
ICE	0.860	0.830	0.939	0.882	0.919	0.782	0.845
ICE+Content	0.887	0.831	0.990	0.903	0.946	0.805	0.870

V. EXPERIMENTS

In this section, we conduct empirical experiments to demonstrate the effectiveness of the ICE model on the Sina Weibo data set. We first introduce settings of our experiments. Then, we compare the ICE model to the state-of-the-art baseline methods. We also study the performance of the ICE model with varying parameters and under different situations. Finally, we analyze the scalability of the ICE model.

A. Experimental Settings

First, we split our data set into training set, testing set, and validation set. Randomly, we use 60% of the events (rumors or non-rumors) in the data set for training, 30% for testing, and the remaining 10% data as the validation set for tuning parameters, i.e., the dimensionality of latent representations and the regularization parameter.

Moreover, we have several evaluation metrics for our experiments: **Accuracy**, **Precision**, **Recall**, and **F1-score**. *Accuracy* is a standard metric for classification tasks, which is evaluated by the percentage of correctly predicted rumors and non-rumors. *Precision*, *Recall* and *F1-score* are widely-used metrics for classification tasks, which are computed according to where correctly predicted rumors or non-rumors appear in the predicted list. The larger the values of above evaluation metrics, the better the performance.

Three competitive methods and their extensions are compared in our experiments:

- News Credibility Propagation (NewsCP) [14] studies how to aggregate credibility from microblogs to events based on a graph optimization method. The classifier we use for each microblog is the widely-used Support Vector Machine (SVM), which is implemented via libSVM⁷ [4]. There are three different versions of NewsCP: **NewsCP-Content**, **NewsCP-Social**, and **NewsCP**. NewsCP-Content only uses the content of microblogs as features. Meanwhile, NewsCP-Social only uses social features of the corresponding microblogs. NewsCP takes usage of all the features. Social features include number of user followers, number of user followees, number of user microblogs, gender of user, user verified or not, number of repostings, number of comments and time of posting.
- The Enquiry Post (EP) model [42] is proposed mainly based on signal tweets. We use our suspicion word

list to identify signal tweets and then apply libSVM [4] for information credibility evaluation. The features used here include percentage of signal tweets, content length, average number of repostings, average number of URLs, average number of hashtags, average number of usernames mentioned, and average time of posting. Considering that **EP** does not take content information of microblogs into consideration, we further make a fusion of EP and NewsCP-Content at the score level and achieve an extended version **EP+Content**.

- Our proposed **ICE** model uses representation learning method to evaluate credibility. Similarly to the EP model, considering ICE only models user behaviors, we make a fusion of ICE and NewsCP-Content at the score level to incorporate content information and achieve an extended version **ICE+Content**.

Note that, the score level fusion means the final predicted score is the sum of the predicted scores of the two methods. Mathematically, for fusing the scores of ICE and NewsCP-Content to generate the score of ICE+Content, it can be calculated as:

$$S_{ICE+Content} = \mu S_{ICE} + (1 - \mu) S_{NewsCP-Content},$$

where $S_{ICE+Content}$, S_{ICE} , and $S_{NewsCP-Content}$ denote predicted credibility scores of methods ICE+Content, ICE, and NewsCP-Content respectively, and μ is selected to be $\mu = 0.8$ in our experiments. For generating the predicted credibility score of EP+Content, the process is the same.

B. Performance Comparison

To investigate the performance of ICE and compared methods, we conduct experiments on the Weibo data set, and report the *Accuracy*, *Precision*, *Precision*, *F1-score*, and *Precision-Recall* curves of these methods.

Table II illustrates the performance comparison with dimensionality $d = 8$ on the Sina Weibo data set evaluated by *Accuracy*, *Precision*, *Precision* and *F1-score*. Using part of the features, NewsCP-Content and NewsCP-Social have the lowest *Accuracy* and *F1-score* among all the methods. Meanwhile, we can see that the performance of NewsCP-Social is better than that of NewsCP-Content. This may indicate that social features are more important than content features for evaluating credibility. Involving both kinds of features, NewsCP achieves great improvement and has a satisfactory performance. Then, EP further improves the performance

⁷<http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html>

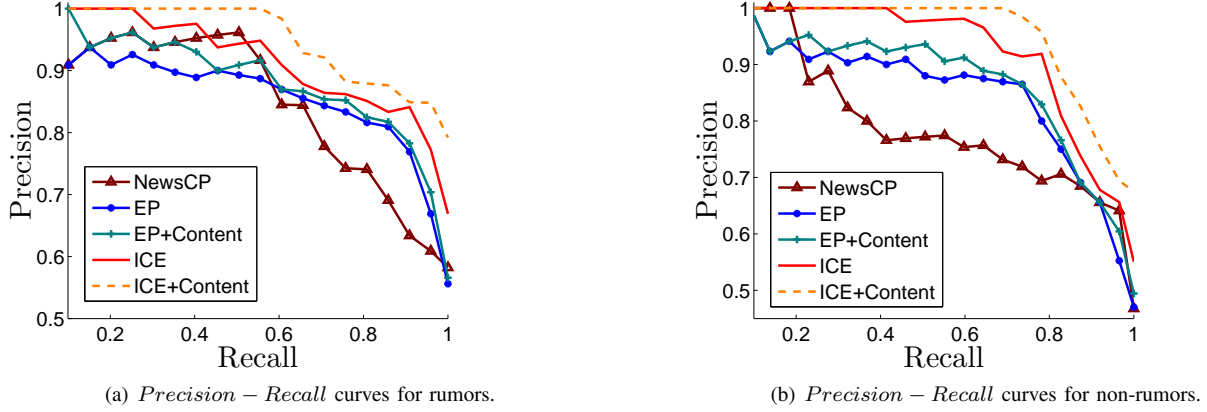
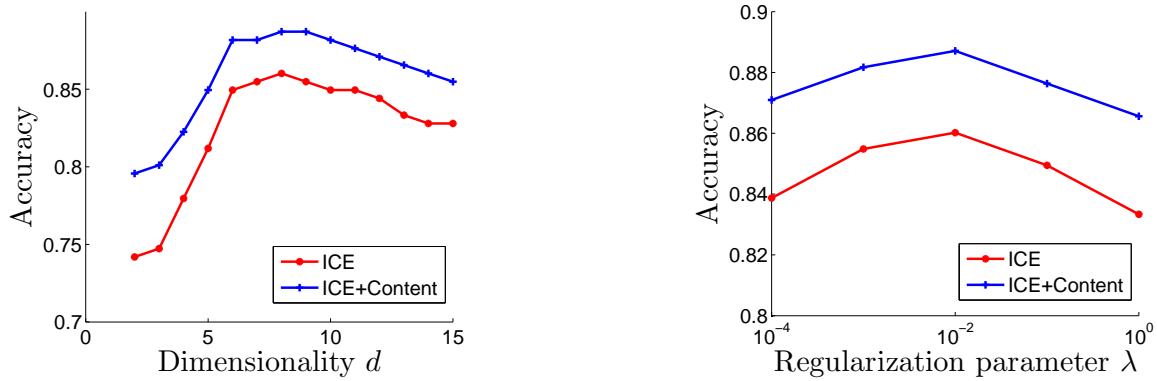


Fig. 6. *Precision-Recall* curves of different methods with dimensionality $d = 8$.



(a) *Accuracy* curves of ICE with varying dimensionality d with $\lambda = 0.01$. (b) *Accuracy* curves of ICE with varying regularization parameter λ with $d = 8$.

Fig. 7. Performance of ICE with varying parameters evaluated by *Accuracy*.

compared to NewsCP and achieves *Accuracy* more than 80%. Incorporating content information of NewsCP-Content, EP+Content achieves a little improvement and becomes the best one among all the compared methods. We can clearly observe that, our proposed ICE model outperforms the compared methods. Incorporating content features, ICE+Content achieves the best performance among all the methods. Compared to EP+Content, ICE and ICE+Content improve the *Accuracy* by 3.7% and 6.4%, respectively. When the target class is rumor, the F1-score improvements are 3.0% and 5.1%, respectively, and when the target class is non-rumor, the improvements are 3.5% and 6.0%, respectively. Moreover, among the results of all the methods, the F1-score of rumors is higher than the F1-score of non-rumors. This may mean that it is more difficult to distinguish non-rumors from rumors than to distinguish rumors from non-rumors.

We also illustrate the *Precision-Recall* curves of the different methods in Figure 6. The *Precision-Recall* curve for rumors in Figure 6(a) shows that ICE+Content outperforms the other methods. The *Precision* of ICE+Content stays 100% until *Recall* is more than 50%. ICE is better than the other compared methods in most cases, except at around *recall* = 45%. The *Precision-Recall* curve for

non-rumors in Figure 6(b) shows that the performance of ICE and ICE+Content is clearly better than that of other methods. The *Precision* of ICE stays 100% until its *Recall* is more than 40%. The *Precision* of ICE+Content stays 100% until its *Recall* is more than 70%. In our experiments, both experimental results in Table II and *Precision-Recall* curves in Figure 6 clearly show that our proposed ICE model achieves satisfactory performances and outperforms the other state-of-the-art methods.

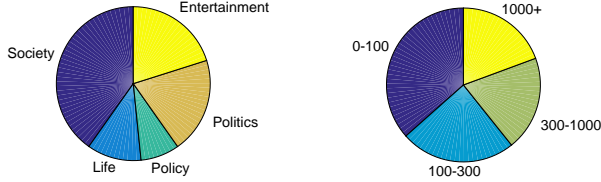
C. Impact of Parameters

To investigate the impact of parameters on the performance of ICE, we illustrate the *Accuracy* performance of ICE with varying parameters in Figure 7. Based on the figure, we can select the best parameters for ICE.

In Figure 7(a), we illustrate the performance of ICE and ICE+Content with varying dimensionality d , where the regularization parameter is set to be $\lambda = 0.01$. The performance of ICE increases rapidly from $d = 3$ and then becomes stable since $d = 6$. ICE achieves the best performance at $d = 8$ and then decreases slightly with increasing dimensionality. Meanwhile, the performance of ICE+Content has the similar trend. However, the *Accuracy* curve of ICE+Content is more

TABLE III
PERFORMANCE COMPARISON UNDER DIFFERENT EVENT TOPICS AND EVENT POPULARITY EVALUATED BY *Accuracy*.

Methods	Topics					Popularity			
	Society	Life	Policy	Politics	Entertainment	0-100	100-300	300-1000	1000+
NewsCP-Content	0.608	0.524	0.667	0.541	0.676	0.662	0.556	0.568	0.611
NewsCP-Social	0.649	0.571	0.667	0.568	0.595	0.647	0.556	0.595	0.639
NewsCP	0.797	0.619	0.933	0.703	0.730	0.794	0.689	0.757	0.778
EP	0.811	0.857	0.800	0.811	0.811	0.824	0.822	0.811	0.778
EP+Content	0.811	0.857	0.867	0.838	0.838	0.853	0.844	0.811	0.778
ICE	0.838	0.905	0.933	0.865	0.838	0.868	0.867	0.865	0.861
ICE+Content	0.851	0.952	0.933	0.946	0.838	0.882	0.933	0.865	0.861



(a) Distribution of different topics. (b) Distribution of different levels of popularity.

Fig. 8. Distribution of different situations in the Weibo dataset.

stable than that of ICE, because the performance of modeling content information is not influenced by dimensionality. From the observation of the curves, we select the best dimensionality of ICE as $d = 8$. Moreover, the curves show that ICE is not very sensitive to the dimensionality in a large range, and ICE still outperforms compared methods even not with the best dimensionality.

The *Accuracy* curves of ICE and ICE+Content with varying regularization parameter λ are shown in Figure 7(b). The performance of ICE grows slowly from $\lambda = 0.0001$ and achieves the highest *Accuracy* at $\lambda = 0.01$. It is obvious that the best parameter for ICE is $\lambda = 0.01$. We can also clearly observe that ICE stays stable in the range of λ from 0.0001 to 1. Moreover, recall results in Table II, even not the best one, the performances of ICE are still better than those of the compared methods.

D. Performance Under Different Situations

We have shown that the proposed ICE model can outperform the state-of-the-art methods. Here, we are going to investigate if ICE can perform better than the compared methods in some specific kinds of situations. We partition our data set according to topics and popularity, and the results evaluated by *Accuracy* under different situations are shown in Table III. The distribution of different situations is shown in Figure 8.

According to topics of the events, we first partition the data set into five categories: society, life, policy, politics and entertainment, as shown in Figure 8(a). Topic “society” talks about all kinds of stuff happening around us, topic “life” contains life skills such as health tips, topic “policy” denotes news about newly released policies, topic “politics” talks about politics,

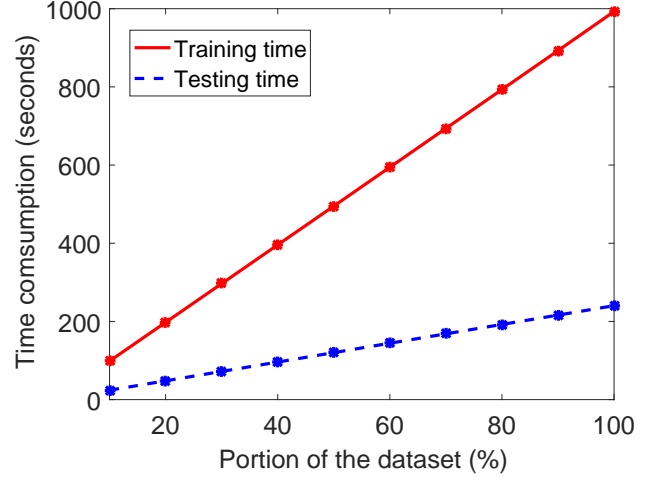


Fig. 9. Training and testing time consumption of ICE with varying portion of the whole Weibo dataset.

government and military, and topic “entertainment” means news about movies, music and sports. Table III illustrates the *Accuracy* of the different methods on the five topics. The results show that our proposed ICE model outperforms the compared methods on all the five topics, and ICE+Content achieves the best performance in all the situations. Meanwhile, NewsCP performs well on the topic of “policy”, and EP has a good performance on the topic of “entertainment”. Moreover, in average, these methods achieve slightly poor performances on topics of “society” and “entertainment” compared to the other topics. This may indicates that news about “society” and “entertainment” has more noise and such rumors are difficult to be identified.

Then, we partition the data set according to the popularity of events. The popularity of an event is computed as the amount of its microblogs, including postings and repostings. As shown in Figure 8(b), the whole data set is partitioned into four categories: 0-100, 100-300, 300-1000, and 1000+. From Table III, we can clearly observe that with all kinds of popularity, ICE outperforms the compared methods, and ICE+Content achieves the best performance. Moreover, we can observe that the larger the popularity, the lower the accuracy in most cases. It may be because that the majority of microblogs contain noise and do not contribute to the evaluation very much. Among the massive microblogs on social media, several significant microblogs are easily hidden by a large amount of

noise. Thus, in the future work, we need to find a method to select significant microblogs, instead of average calculation.

E. Scalability Analysis

Besides the analysis of the effectiveness of ICE, we also investigate the scalability of the ICE model with varying portions of the Weibo data set. The model is implemented with Python⁸ and Theano⁹. The code is run on a computer with 4 Core 2.5 GHz CPU and 16 GB RAM, and the GPU model is NVIDIA Tesla K20Xm. On the Weibo data set, we measure the corresponding time cost of one iteration in both training and testing process. Figure 9 shows the time consumption with varying portions of the whole data set. We can observe that both training and testing time consumption of ICE are linear with respect to the size of data set. This shows the scalability of ICE. Our proposed model not only can achieve the state-of-the-art prediction performance but also can run effectively on large-scale data.

VI. SYSTEM

As introduced and discussed above, we have achieved a information credibility evaluation model ICE with state-of-the-art performances. Rather than only using the ICE model in academic data sets and research, it is vital to construct a real-time information credibility evaluation system on social media and make our proposed model applied in real applications. Thus, based on our model and the Sina Weibo data set, we built a Network Information Credibility Evaluation (NICE) system [35]. NICE is a webpage-based system that can automatically crawl online information from Sina Weibo and evaluate the credibility of online information that users enquire.

Figure 10 illustrates the flow chart of the NICE system. Using the system, a user can input a query to retrieve the related information. If a user's query matches rumors in the Weibo data set, users can identify the rumor immediately. Otherwise, NICE will crawl real-time information from social media, i.e., Sina Weibo. Then, the user can select one microblog to evaluate the corresponding credibility based on our model. Based on the selected microblog, the system will crawl all the related microblogs from Weibo and collect related information including content information, temporal information, comment information, and corresponding user profiles. Based on this information and the trained model, NICE can evaluate the credibility of the related information and provide a predicted score of the event. With our proposed ICE model, the NICE system can achieve great performances in information credibility evaluation. It can be applied effectively and stably in information credibility evaluation and online information management on social media.

VII. CONCLUSIONS AND FUTURE WORK

In this work, to evaluate information credibility on social media, a novel method, i.e., ICE, has been proposed. ICE aims to learn dynamic representations for the microblogs that

describe events spreading on social media. The learning is based on user credibility, behavior types, temporal properties, and comment attitudes. The aggregation of these key factors makes the dynamic and joint representations of microblogs, and the aggregation of representations of all the microblogs during information spreading can generate the credibility representation of events on social media. Experiments conducted on a real data set crawled from Sina Weibo show that ICE outperforms the state-of-the-art methods.

In the future, we can further investigate the following directions. First, in ICE, the content information has not been considered when learning credibility representations. We plan to analyze the event content and extract main elements in it to predict the happening probability of the event based on a large news database. Second, information about an event on other platforms, e.g., news websites and forums, can be incorporated in our model. Third, for the aggregation of microblogs of an event, we use average computation in ICE, which is clearly not the best solution. We need to find a method to select significant microblogs.

REFERENCES

- [1] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828, 2013.
- [2] S. Bourigault, C. Lagnier, S. Lamprier, L. Denoyer, and P. Gallinari. Learning social network embeddings for predicting information diffusion. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 393–402. ACM, 2014.
- [3] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.
- [4] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.
- [5] N. DiFonzo and P. Bordia. Rumor, gossip and urban legends. *Diogenes*, 54(1):19–35, 2007.
- [6] A. M. Elkahky, Y. Song, and X. He. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *Proceedings of the 24th International Conference on World Wide Web*, pages 278–288. International World Wide Web Conferences Steering Committee, 2015.
- [7] S. Feng, X. Li, Y. Zeng, G. Cong, Y. M. Chee, and Q. Yuan. Personalized ranking metric embedding for next new poi recommendation. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 2069–2075. AAAI Press, 2015.
- [8] K. D. Giudice. Crowdsourcing credibility: The impact of audience feedback on web page credibility. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–9, 2010.
- [9] A. Grover and J. Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining*, pages 855–864. ACM, 2016.
- [10] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 729–736. International World Wide Web Conferences Steering Committee, 2013.
- [11] M. Gupta, P. Zhao, and J. Han. Evaluating event credibility on twitter. In *Proceedings of the 12th SIAM International Conference on Data Mining*, pages 153–164. SIAM, 2012.
- [12] Y. Jacob, L. Denoyer, and P. Gallinari. Learning latent representations of nodes for classifying in heterogeneous social networks. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 373–382. ACM, 2014.
- [13] F. Jin, W. Wang, L. Zhao, E. Dougherty, Y. Cao, C.-T. Lu, and N. Ramakrishnan. Misinformation propagation in the age of twitter. *Computer*, 47(12):90–94, 2014.

⁸<https://www.python.org/>.

⁹<http://deeplearning.net/software/theano/>.

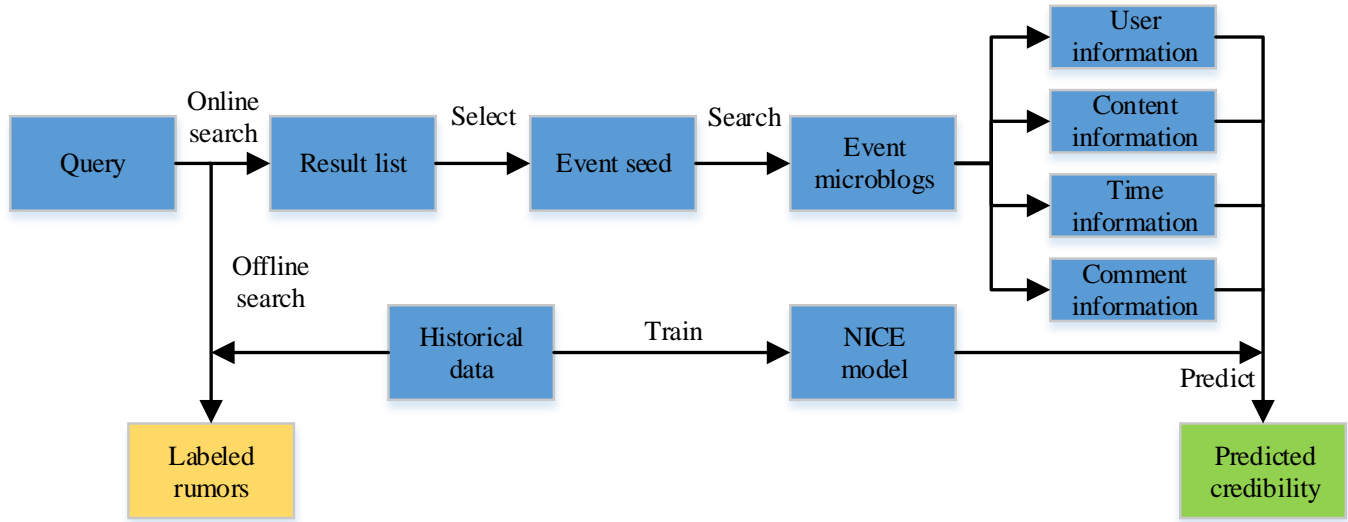


Fig. 10. Overview of the Network Information Credibility Evaluation (NICE) system.

- [14] Z. Jin, J. Cao, Y.-G. Jiang, and Y. Zhang. News credibility evaluation on microblog with a hierarchical propagation model. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 230–239. IEEE, 2014.
- [15] Z. Jin, J. Cao, Y. Zhang, and J. Luo. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 2972–2978. AAAI Press, 2016.
- [16] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang. Prominent features of rumor propagation in online social media. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 1103–1108. IEEE, 2013.
- [17] X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava. Truth finding on the deep web: Is the problem solved? *Proceedings of the VLDB Endowment*, 6(2):97–108, 2012.
- [18] Q. Liu, S. Wu, and L. Wang. Collaborative prediction for multi-entity interaction with hierarchical representation. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pages 613–622. ACM, 2015.
- [19] Q. Liu, S. Wu, and L. Wang. Cot: Contextual operating tensor for context-aware recommender systems. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 203–209. AAAI Press, 2015.
- [20] Q. Liu, S. Wu, L. Wang, and T. Tan. Predicting the next location: A recurrent model with spatial and temporal contexts. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 107–113. AAAI Press, 2016.
- [21] Q. Liu, F. Yu, S. Wu, and L. Wang. A convolutional click prediction model. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pages 1743–1746. ACM, 2015.
- [22] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pages 1751–1754. ACM, 2015.
- [23] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur. Recurrent neural network based language model. In *INTERSPEECH*, volume 2, page 3, 2010.
- [24] T. Mikolov, S. Kombrink, L. Burget, J. H. Černocký, and S. Khudanpur. Extensions of recurrent neural network language model. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5528–5531. IEEE, 2011.
- [25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
- [26] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.
- [27] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599. Association for Computational Linguistics, 2011.
- [28] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 452–461. AUAI Press, 2009.
- [29] S. Y. Rieh, G. Y. Jeon, J. Y. Yang, and C. Lampe. Audience-aware credibility: From understanding audience to establishing credible blogs. In *Proceedings of the 8th International Conference on Weblogs and Social Media*. AAAI Press, 2014.
- [30] Y. Shi, A. Karatzoglou, L. Baltrunas, M. Larson, and A. Hanjalic. Cars2: Learning context-aware representations for context-aware recommendations. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 291–300. ACM, 2014.
- [31] S. Sun, H. Liu, J. He, and X. Du. Detecting event rumors on sina weibo automatically. In *Web Technologies and Applications*, pages 120–131. Springer, 2013.
- [32] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077. International World Wide Web Conferences Steering Committee, 2015.
- [33] D. A. Waguih and L. Berti-Equille. Truth discovery algorithms: An experimental evaluation. *arXiv preprint arXiv:1409.6428*, 2014.
- [34] P. Wang, J. Guo, Y. Lan, J. Xu, S. Wan, and X. Cheng. Learning hierarchical representation model for next basket recommendation. In *Proceedings of the 38th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 403–412. ACM, 2015.
- [35] S. Wu, Q. Liu, Y. Liu, L. Wang, and T. Tan. Information credibility evaluation on social media. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 4403–4404. AAAI Press, 2016.
- [36] S. Wu, Q. Liu, L. Wang, and T. Tan. Contextual operation for recommender systems. *Knowledge and Data Engineering, IEEE Transactions on*, 28(8):2000–2012, 2016.
- [37] Y. Wu, C. DuBois, A. X. Zheng, and M. Ester. Collaborative denoising auto-encoders for top-n recommender systems. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 153–162. ACM, 2016.
- [38] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. *Knowledge and Data Engineering, IEEE Transactions on*, 20(6):796–808, 2008.
- [39] X. Yin and W. Tan. Semi-supervised truth discovery. In *Proceedings of*

- the 20th international conference on World wide web*, pages 217–226. ACM, 2011.
- [40] F. Yu, Q. Liu, S. Wu, L. Wang, and T. Tan. A dynamic recurrent model for next basket recommendation. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 729–732. ACM, 2016.
- [41] W. Zhang, T. Du, and J. Wang. Deep learning over multi-field categorical data - - A case study on user response prediction. In *Proceedings of the 38th European Conference on Information Retrieval*, pages 45–57, 2016.
- [42] Z. Zhao, P. Resnick, and Q. Mei. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1395–1405. International World Wide Web Conferences Steering Committee, 2015.